专利无效对比文件判定方法研究*

■ 郭诗琪[1,2]　贠强[2]　陈亮[2]　周杰[2]
[1] 中国医学科学院医学信息研究所　北京 100020　[2] 中国科学技术信息研究所　北京 100038

摘　要：[目的/意义] 对比文件是用以判断专利能否授权或无效的重要文件，针对传统信息检索方法的不足且鲜有利用机器学习方法研究对比文件检索的问题，在引入对比文件信息的基础上，构建专利相关性判定模型。[方法/过程] 以专利无效判决书中的目标专利与对比文件为数据集进行实验，提取文本相似度、共现词汇和共词数量特征信息，利用 GBDT 模型将对比文件的检索问题转化为判断其是否相关的分类问题。[结果/结论] 研究结果表明，不同字段数据对分类效果的贡献不同，其中说明书字段的准确率、召回率和 F1 值分别为 79%、48% 和 59%，并且多特征集成后的分类效果显著优于单一文本相似度的结果，最后对实验错分情况进行分析，指出本研究下一步的研究方向。

关键词：专利无效宣告　对比文件　特征选择　机器学习
分类号：TP181
DOI：10.13266/j.issn.0252-3116.2021.02.012

随着经济全球化进程加快，科技创新在国民经济发展中的驱动作用不断加强，各国对技术创新的知识产权保护愈发重视，其结果是专利数量爆发性增长。以中国为例，2018 年国家知识产权局受理的国内外专利申请量较 1985 年增长近 300 倍[1]，相比之下专利审查工作目前仍然以"检索系统 + 人工判读"为主，成本高、效率低、受审查员自身专业背景和技术水平等主观因素影响，其不仅导致大量待审专利申请的积压[2]，而且极易出现审查漏洞并导致专利的错误授权，为技术持有人的后续市场行为带来了严重风险[3-4]。在目前对专利审查质量和审查效率要求愈发严格的大环境下[5-6]，这种矛盾更加突出。因此如何有效提升专利审查的质量和效率，成为一个摆在知识产权管理部门和从业者面前亟待解决的重要问题。

在影响专利审查质量和效率的各种因素中，对比文件判定是其中的关键因素和主要瓶颈。所谓对比文件，即用来判断发明或实用新型是否具备新颖性、创造性等所引用的相关文件[7]。对比文件的判定能力一直以来都是反映专利审查员和相关从业者水平高低的重要标记，通过访谈第三届专利检索大赛[8]优胜者得知，即便国内顶尖的专利审查员，要在 4 个小时内获取一篇有效的对比文献，也是件非常困难的事情，其难点不仅在于专利包含了丰富和高度专业化的技术内容、法律内容，更在于这些内容由于商业、技术等方面的考量，经常会辅以文字变换、上下位概念替换、惯用技术手段置换以及对部分内容以隐式方式加以公开[9]，从而使普通基于倒排索引和文本相似度计算的专利检索系统难以应对。

然而随着第三次人工智能浪潮的到来，以统计机器学习为代表的人工智能技术为对比文献判定的自动化提供可能性。这也构成了本文的研究主题：跳出将对比文件判定作为信息检索问题的传统思路，以一种有监督学习的方式对对比文件进行识别判定，具体来说，就是首先获取目标专利及其对应对比文件的金标准，进而创建训练数据集并从目标专利和候选文献的关联信息中提取出多种特征，最终以分类方式识别隐藏在候选文献中的真实对比文件。相比信息检索方法，新方法的优势不仅在于超越文本相似度，将更多能

* 本文系国家重点研发计划项目课题"知识产权信息智能采集及深加工技术研究与应用示范"（项目编号：2017YFB1401902）和中信所重点工作"重点科技领域前沿跟踪与深度研究"（项目编号：ZD2020-02）研究成果之一。

作者简介：郭诗琪（ORCID：0000-0002-9311-8088），硕士研究生；贠强（ORCID：0000-0002-9156-6063），副研究员，博士，硕士生导师，E-mail：yunq@istic.ac.cn；陈亮（ORCID：0000-0002-3235-9806），副研究员，博士，硕士生导师；周杰（ORCID：0000-0003-0147-8674），通讯作者，信息资源中心副主任，研究馆员，硕士生导师。

收稿日期：2020-03-10　修回日期：2020-10-12　本文起止页码：117-125　本文责任编辑：易飞

够有效提升对比文件判定的特征引入进来,更在于可以通过错误分析,获取当前方法在数据处理、特征工程和模型构建上的不足之处,为未来的方法优化指明道路。

1 相关研究

1.1 对比文件的基本内涵

对比文件包括专利文件和非专利文件[7],通常将待判断的发明或实用新型称为目标专利。

围绕专利生命周期的各个时期,根据检索目的的不同,可以将专利检索分为现有技术状况检索、无效性检索、侵权检索、确权检索等。其中无效性检索是因无效请求人对专利权产生质疑而发起的[10],目的是检索因审查疏漏或对现有技术隐瞒而造成的错误授权的证据[11],进而对发明产生时的新颖性进行复审,其中无效证据的查找是无效请求能够成功的关键所在[12]。

根据检索报告中对比文件与权利要求的关系[7]可以将对比文件分为 X、Y、A、R、P、E 几类,其含义如表 1 所示,其中 X 和 Y 类均与目标专利密切相关。

表 1 对比文件的类型及含义

类型	定义
X	单独影响权利要求的新颖性或创造性的文件
Y	与检索报告中其他 Y 类文件组合后影响权利要求的创造性的文件
A	背景技术文件,即反映权利要求的部分技术特征或者有关的现有技术的文件
R	任何单位或个人在申请日向专利局提交的、属于同样的发明创造的专利或专利申请文件
P	中间文件,其公开日在申请的申请日与所要求的优先权日之间的文件,或者会导致需要核实该申请优先权的文件
E	单独影响权利要求新颖性的抵触申请文件

1.2 对比文件的检索

1.2.1 传统的对比文件检索

审查员或无效请求人一般通过与目标专利相关的技术关键词[10]或组配专利分类号[13]在海量数据中进行检索,虽然组配的方式在一定程度上提高了检索效率,但仍存在歧义词、同义词等影响因素。即使是经验丰富的审查人员利用深入挖掘发明人/申请人信息、追踪相关文献的参考文献信息、追踪前沿领域的原创性非专利文献等[14]高效获取对比文件的追踪检索[15-16]的技巧,也难以避免反复多次构建检索式、阅读理解相关文献并判断其能否作为对比文件的繁琐过程。

K. Rajshekhar 等证明了目标专利与至少 20% 的高度相关专利之间没有明显相似的技术术语,最先进的语义检索技术也最多只能检索出高度相关的现有技术的十分之一[17]。隆瑾基于专利引文和对比文件都与目标专利的技术内容存在一定的相似性的特点,提出从专利引文中获取对比文件的思路[18],虽然在理论层面具有较高的指导价值,但是对于实践来说缺乏一套具体的检索方法。现有的关于对比文件的检索思路对检索经验的依赖较高。

1.2.2 机器学习在对比文件检索中的应用

传统的信息检索模型对检索主题和待检索文档的相关性进行排序,主要利用词频、逆文档频率和文档长度这几个因素来人工拟合排序公式,根据排序返回查询结果。

随着相关度的影响因素变多,基于大数据的学习排序(learning to rank)逐渐成为热门领域。学习排序可以把各个现有排序模型的输出作为特征,然后训练一个新的模型,并自动学习这个新的模型的参数。简单地说,学习排序是组合多个现有的排序模型来生成新的排序模型的算法。利用机器学习技术来对搜索结果进行排序是近几年热门的研究领域,但是该算法在无效专利检索的领域并未见相关应用和研究。

国内研究多从相似专利识别角度切入,张杰等提出了利用权利要求书文本的主谓宾结构进行相似专利的识别[19]。刘玉琴等在二步检索的基础上结合中文专利独立权利要求结构的特征信息构建中文专利无效检索模型[20]。传统方法多利用文本相似度或检索系统,如 cosine similarity、elastic Search 等进行对比文件识别,但因其只进行了字符串间的比对而导致实验效果有限。而基于权利要求书文本结构独特性的分析方法多从专利文本的分词模型[21]、命名实体识别、文本分类[22-24]等角度切入,在此基础上可以进一步探索相关专利的识别与检索问题。

国外学者在这方面的研究起步略早,大多采用机器学习的方法,包括利用加权最大置信度方法对仅利用词频的专利特征挖掘[25]方法进行优化、基于专利元数据和引文信息的专利主题自动分类[26]、协同训练方法标注摘要中功能字句[27]等,从而探索专利检索精度的提高。

F. Kreuchauff 等利用服务机器人领域的小型核心专利数据集中的标题、摘要和 IPC 信息,提出基于词性、引文或联合方法的专利检索策略[28]。伯克利分校的 W. Ho 等利用机器学习技术开发了一个基于文本相似度的预测 PTAB 受理专利无效请求概率的程序[29-32]。斯坦福大学的 L. Ryan 等利用授权专利占公司申请专利数量的比重、专利审查员审查通过率等元

数据特征和卷积神经网络进行专利授权预测，并证实了该方法效果优于仅利用专利文本数据的模型[33]。

机器学习技术的辅助确实优于传统的检索方法，国内外基于专利文本、结构化信息、题录信息等的研究已经较为完备，此外亟需改进术语间相关性判断方法、语义表示方法或融入专家经验以进一步提升相关文件的检索效果[34]。本研究重点关注无效性检索环节中的对比文件获取问题，利用机器学习方法探索从无效证据数据库中识别并获取对比文件的问题。

2 对比文件自动识别的研究设计

2.1 总体框架

本文以将对比文件的检索问题转化为机器学习中判断对比文件与目标专利是否相关的 0－1 分类问题为研究思路，纳入除文本相似度外更丰富的特征进而实现对比文件的识别，该研究总体框架如图 1 所示，按照“数据集构建－数据预处理－特征选择与提取－标签信息提取－模型测试”的思路展开，探索利用机器学习方法解决人工检索问题的可行性。详细步骤见表 2。

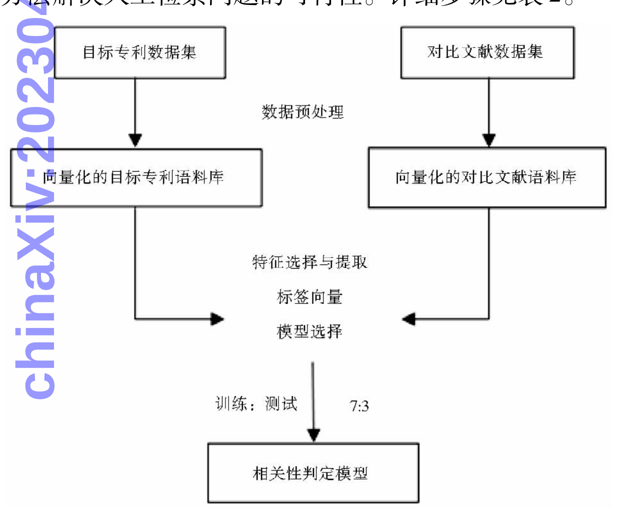


图 1 专利相关性判定模型技术路线

表 2 研究基本步骤

步骤	内容
1	根据专利无效判决书，构建目标专利数据集和对比文献数据集
2	数据清洗、规范化并生成向量化的目标专利和对比文献语料库
3	特征选择、提取与规范化，形成特征数据集
4	根据专利申请号间的匹配关系，生成表征数据集间实际对应关系的标签向量
5	模型测试与结果评估

2.2 数据准备

鉴于发明专利具有保护范围更广泛、稳定性更优、法律保护强度更大的优势[35]，本研究将数据范围界定为向专利复审委员会提出无效宣告请求的中国发明专利。

本研究将万象云数据库作为数据来源，首先下载 1990－2018 年间经历无效审查的专利作为目标专利样本总体，共计 4 246 件，每条样本内容包括无效宣告专利的决定号、无效请求人、专利权人等基本信息以及无效宣告的法律依据、决定要点和决定书全文等字段。

专利无效判决书是由判决依据的法律条款、决定要点、决定全文构成的，其中决定全文中包含以下 4 部分内容：①目标专利基本信息；②无效请求人提出请求的原因、依据条款、证据附件等信息；③证据认定的详细结果；④案件决定，即根据证据与专利法判定该专利的法律状态（维持有效、维持部分有效、全部无效）。

接下来，利用正则表达式从无效判决书中无效请求人提供的专利类型的无效证据文本中提取无效证据专利号，共计 21 718 个；再从万象云数据库中批量检索，下载其专利名称、摘要等题录项作为无效证据样本总体。

最后从目标专利样本总体中随机抽取 60 件专利作为目标专利数据集，并根据专利号从无效证据样本总体中抽取与目标专利相匹配的无效证据作为对比文件数据集，共计 299 件，其中专利名称、摘要、权利要求和说明书字段是最重要的研究数据。本研究所使用到的专利字段及含义如表 3 所示：

表 3 专利字段及其内涵与功能

专利文献字段	专利字段内涵与主要功能
专利名称	简短、准确地表明专利要保护的主题和类型
摘要	写明专利名称和所属技术领域，清楚反映所要解决的技术问题
权利要求书	以说明书为依据，清楚简要地限定要求专利保护的范围。记载发明或实用新型的技术特征，是专利审查的依据
说明书	清楚完整地描述发明或实用新型，使所属技术领域的技术人员能够理解和实施该发明或实用新型。包括技术领域、背景技术、发明内容、附图说明和具体实施方式

注：表 3 中文字描述均来自于《专利审查指南 2010》

2.3 特征提取

该步骤通过选取有价值的关键信息，剔除噪音来使分类器学习到文本中最重要的信息，进而提高分类器性能。在本研究中，我们主要使用文本相似度、共现词汇和共现词汇数量三种特征，具体提取流程如图 2 所示，其中：

（1）文本相似度。首先分别训练 TF-IDF 模型得到用 tf-idf 值表示的专利各字段文档向量，即标题、摘要、权利要求书、说明书与合并文档（本文中将专利标题、

摘要、权利要求书与说明书的合并文本称为合并文档），然后据此训练 LDA 模型，再分别将各字段文档映射到主题空间，最后分别求得各字段间的文本相似度。

（2）共现词汇。文本相似度是一个重要的特征，但是仅使用文本相似度得到的相关性结果较差。目标专利与对比文件间若存在共现词汇，则它们可能相关，且存在的共现词汇越多，它们之间相关的可能性就越大。

取目标专利和对比文件间各个对应字段文档的词汇交集作为共现词汇。提取共现词汇特征时，相比于不加筛选地使用，本研究采用信息增益（information gain）的方案，先筛选信息增益 Top 600（在对比 Top50、100、250、300、600 后选定最优情况 Top600）的词汇形成词典，再将词典中的词汇向量化作为特征。信息增益是指加入特征 X 的信息使得类别 Y 的不确定性减少的程度[36]。利用信息增益度量词汇的重要性不仅可以减少词汇噪音，还可以减少存储和计算负担。

（3）共现词汇数量。目标专利与每个对比文献对应字段的共现词汇数量取值在[0，∞]，为减小方差对模型学习其他特征能力的干扰，调用了 sklearn. preprocessing 中的 MinMaxScaler 标准化方法将贡献词汇数量规范化至[0，1]，得到共现词汇数量特征。

最后将文本相似度特征、共现词汇特征、共现词汇数量特征合并成一个特征矩阵，作为本研究的特征数据集。

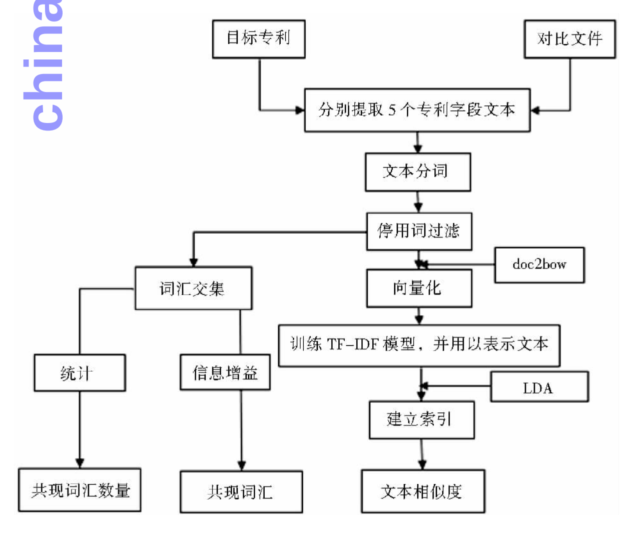


图 2　特征提取流程

2.4　标签生成

该步骤用以产生目标专利和对比文献是否匹配的标签信息，也是实验中判断专利之间关系的金标准。如果某对比文件确实是无效判决书中无效请求人提交的用来证明某目标专利无效的证据，即该对比文件与目标专利之间具有匹配关系，则将该对比文件与目标专利的匹配标签置为 1，否则为 0。用这种方式生成的标签信息作为后续模型分类效果的评判标准。

2.5　模型测试与结果评估

机器学习中的各类算法模型发展迅速并在各个领域发挥着重要的作用，如 logistic 回归模型、隐马尔可夫模型、条件随机场模型等等。本研究采用梯度提升决策树模型（Gradient Boosting Decision Tree，简称 GBDT）应用于相关性判定中，GBDT 是 J. H. Friedman 于 2001 年提出的一种提升算法[37]，主要包括计算候选分裂点、创建决策树、寻找分裂树节点、计算合并叶子节点的预测值几部分。具体来说，首先初始化预测值，然后进入一个迭代过程，每次增加一棵分类树并从新的叶子节点中得到预测值及其与实际值之间的残差，接下来根据残差进行学习，生成新的分类树，循环至实际值与由最终分类树得到的预测值之间的残差足够小。由于该模型性能优良，被广泛应用于数据竞赛和工程实践。

本研究的关键环节在于文本特征和标签向量的提取，利用 Google 开发的 Python 机器学习库 scikit-learn[38]，按照 7∶3 划分训练集和测试集，输入 GBDT 中进行学习，主要使用准确率与召回率的调和平均数 F1 值来反映模型的分类效果。

3　实证研究

3.1　实验概况

本研究主要利用 Gensim[39] 的框架进行实验，实验代码基于 Python2.7.13，主要开发环境为 Windows7，64 位操作系统，处理器为 Intel 的 16 核处理器，运行内存为 64G。整体技术路线见图 3。

3.2　数据获取

我们以万象云数据库[40] 作为数据来源，首先从目标专利样本总体中随机选取 60 件作为目标专利样本，人工下载并收集整理其专利基本信息，包括目标专利申请号、专利名称、专利摘要、权利要求书、说明书等，构建目标专利数据集；然后从无效证据样本总体抽取与目标专利数据集中目标专利号相匹配的无效证据专利号，下载其专利名称、摘要、权利要求、说明书等基本信息，构建对比文献数据集。

经处理后，本研究所需的目标专利数据集和对比文献数据集分别包含 60 件和 299 件专利。

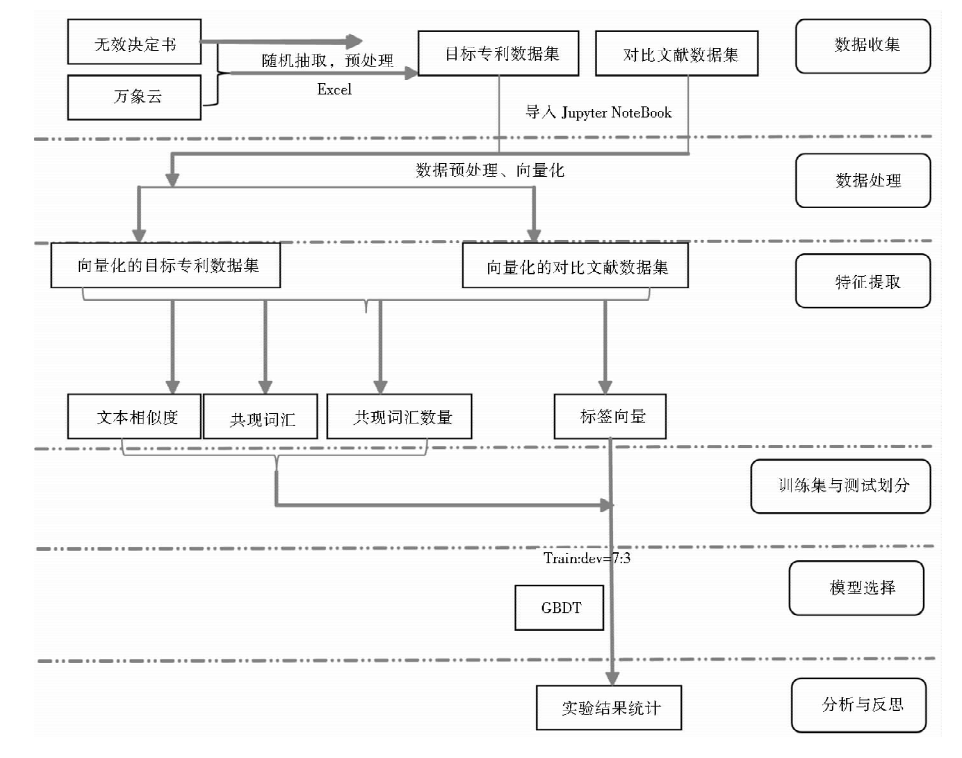


图 3　相关性判定模型研究流程

3.3　实验结果

实验经过多次调整参数,选择参数 n_estimators = 20 为模型性能最好的情况,n_estimators 为超参数,指的是弱学习器的最大迭代次数。实验结果用指标 A、P、R、F1 进行评价,其指标具体含义见表 4。

本次实验主要结果如表 5 所示,表中 Title、Abstract、Claim、Description 和 All 分别代表字段专利标题、摘要、权利要求、说明书和合并文档,对照组表示使用合并文档字段仅利用文本相似度作为分类特征的对照实验。

表 4　GBDT 模型实验评价指标及其含义

指标	含义
准确率 A	是指预测正确的专利对占总的专利对的比重
精确率 P	正确预测为相关的专利对占全部预测为相关的专利对的比重
召回率 R	正确预测为相关的专利对占实际具有相关关系的专利对的比重
F1 值	精确率和召回率的调和平均数,常用该指标做综合评价

表 5　GBDT 分类结果统计

指标	Title	Abstract	Claims	Description	All	对照组
A	0.986 3	0.986 4	0.987 4	0.990 3	0.989 4	0.985 1
P	0.568 2	0.549 3	0.730 8	0.791 7	0.701 8	0.500 0
R	0.312 5	0.487 5	0.237 5	0.475 0	0.500 0	0.062 5
F1	0.403 2	0.516 6	0.358 5	0.593 7	0.583 9	0.111 1

同是使用合并文档字段作为实验数据,加入共现词汇和共现词汇数量作为分类特征的实验组(All)与仅为文本相似度的对照组相比,后者分类效果呈现出断崖式下降,从综合指标 F1 来看,多特征的分类效果几乎是对照组的 5 倍。由此可以认为本实验优于传统的单纯使用文本相似度特征的分类效果。

3.4　实验讨论

本节将在实验结果的基础上,从字段、特征和误差 3 个角度进行讨论,为下一步的研究寻找思路。

3.4.1　字段评估

由图 4 可知,在 GBDT 模型中单独使用各字段时的分类效果(以 F1 值为评价指标)分别为:0.593 7(说明书)＞0.583 9(合并文档)＞0.516 6(摘要)＞0.358 5(权利要求)＞0.403 2(标题)。

从图 4 可以更为直观地看出 GBDT 模型中,说明书是分类效果最优的字段,而集标题、摘要、权利要求、说明书于一体的合并文档字段分类效果与说明书类似,甚至有些许下降。对分类效果较好的几个字段进行组合(如说明书＋摘要;标题＋摘要＋权利要求等多

个方案),实验发现分类效果与表 5 中合并文档的效果近似,这说明文本容量的增多并不是分类效果提升的充分条件,效果最佳的说明书承载了专利最详细的信息,包含技术领域、技术背景、发明内容、具体实施方式等更加丰富和细节的信息,从一定程度上反映出文本内容之间相关性和独特性的丰富度也是分类效果的重要影响因素。

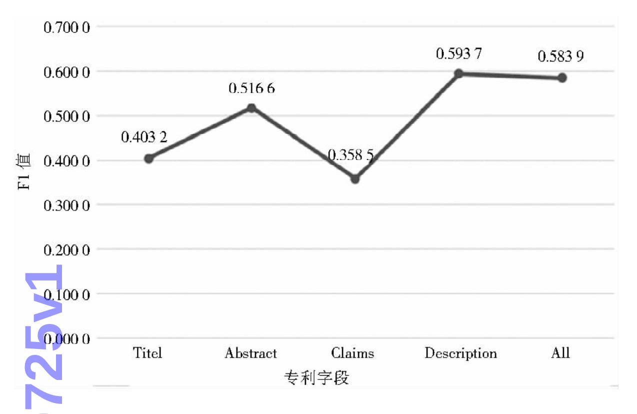


图 4 GBDT 模型各字段最优分类效果

除说明书字段外,文本容量与说明书有较大悬殊的摘要字段也表现出较强的分类能力,而权利要求书分类能力最弱。这是由于摘要是一个专利内容的浓缩,虽然有 300 字的字数限制[41],但是囊括了主要的背景和技术信息,而权利要求书作为专利保护范围的法律依据,会使用大量较为专深和非常用词,除了对技术信息的详细描述外,更多包含的是与其他相关专利在新颖性和创造性方面的差异,包含了该专利的技术细节和独特之处。

3.4.2 特征评估

本实验采用文本相似度、归一化的共现词汇数量以及 600 个共现词汇作为特征,对其特征的权重,即不同特征对实验结果的贡献进行排序,得到图 5,从图中可看出权重最高的基本上都是高度专业化的共现词汇。

本次实验中有 64 个特征的权重大于 0,其范围为[0,0.072],且文本相似度和归一化的共现词汇数量分别排在第 2 和 23 位。由此可见在 GBDT 模型中所有的特征并非均有贡献于模型分类,贡献最大的是高度专业化的共现词汇。多次实验后发现类似的规律,基本上文本相似度和大多数共现词汇特征的权重远大于归一化的共现词汇数量。

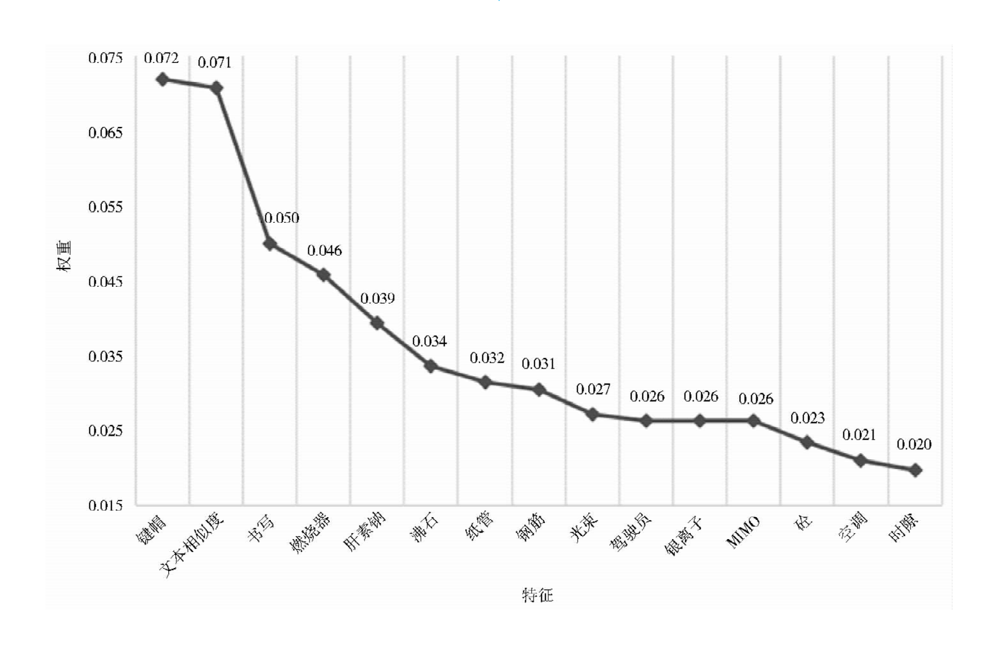


图 5 GBDT 模型权重序前 15 位的特征

3.4.3 误差分析

误差分析指的是检查被算法误分类的开发集样本的过程,是机器学习中常见研究方法,以便于获得探索新方向的灵感。本节在专利文本的基础上,引入专家干预,从专利之间的语义相关性、领域相关性等角度进行了分词误差、参数影响、领域、机器翻译等方面的分析,以便获知各个误差原因及优化的优先级。本次实验测试集中出现误差的专利对共计 52 件,分为实为对比文件而判定不相关(FN)和实际不相关而判定为对比文件(FP)两种情况,分别占 42 和 10 件。

从文本相似度角度进行分析,发现 FN 和 FP 的专利文本相似度区间分别为[0.096,0.995]和[0.041,

郭诗琪，贠强，陈亮，等. 专利无效对比文件判定方法研究[J]. 图书情报工作,2021,65(2):117-125.

0.563]；从共现词汇角度发现，误分类的专利文本中出现最多的共现词是“砼”（0.023 437）和“水”（0.000 588）。FN 和 FP 情况下的误差数量和比例如表6 所示：

表 6　误差分析

误差原因		参数	分词	语义相关	文本不相关	领域	机器翻译
误差含义		生物、化学等领域的专利文本分词后出现常见的数值、单位或术语缩写等问题（超过 5 处）	存在超过 5 处明显的分词错误（参数原因除外）	文本相似度极低但领域专家判定其语义具有相关性的情况	人工阅读后发现文本相关度较低	领域专家判定两件专利的领域是否相同	存在专利文本不流畅的情况（仅见于外国专利）
FN	数量(个)	14	37	7	9	25	14
	比例(%)	33.33	88.10	16.67	21.43	59.52	33.33
FP	数量(个)	5	6	0	1	5	4
	比例(%)	50.00	60.00	0.00	10.00	50.00	40.00

根据上述错误分析结果，本文认为可以从两个方面尝试进行改进：①在数据准备环节区分专利领域，因为领域相同的专利之间更有可能具有相似的技术背景等文本和语义信息，所以同领域专利之间更可能对彼此的稳定性产生干扰，而不同领域的专利之间影响稳定性的可能性较低，比如生物领域的专利极少出现被机械领域的专利证明无效的情况；②尽可能提高数据处理精度，如分词精度、停用词过滤精度等，尽可能将可控的误差降到最低。总体而言，本实验得出的结论与优化建议如表 7 所示：

表 7　实验结论

结论	具体内容	展望
字段	说明书是承载专利信息最重要的文本	未来实验可以保留说明书和摘要字段的数据以加快实验进程
特征	按照贡献大小：高度专业化的共现词汇 > 文本相似度 > 贡献词汇数量	添加领域相似度作为模型特征
误差	1. 从文本角度来看，分词误差和文本中的参数数值是最重要的优化点；2. 从特征角度来看，领域相似性是重要的分类特征	提高数据预处理精度

4　结语

无效专利对比文件查找中仍然使用传统的信息检索方法，例如文本相似度、倒排序索引等，但专利文本中存在诸多特殊之处，比如大量同义词、对等词、概念泛化等现象，这些问题会导致利用传统方法检索对比文件效果不佳的情况。对此，本文利用机器学习方法将繁琐而低效率的对比文件检索问题转化成了判断目标专利与对比文献是否相关的分类问题，在此基础上将单一的相似度特征扩展到文本相似度、共现词汇以及共现词汇数量 3 个特征信息进行专利相关性判定。实验结果证实了本实验方法的有效性和可行性。

本研究的主要贡献包括以下两方面：①除了文本相似度这种从篇章级别判断文本相关性的特征，本研究更将特征下探到词汇级别，对文本相关性进行判断，这其中使用的特征包括文本相似度、共现词汇和共现词汇数量；②利用机器学习中的分类方法代替传统的信息检索方法来对文本相关性进行判断，使用 GBDT 模型并取得了良好效果；③同时做出了详细的误差分析，这有助于指出实验中误分类的原因和类型，提供下一步改进的方向。

然而，本研究也存在不足之处：①对比文件查找分为两个步骤，第一步是检索相关文件，第二步是判定检索结果中的对比文件，也是本研究的重点所在，有关专利检索的相关研究将是下一阶段的工作。②目前研究基于小数据集，如何扩充到大数据集并解决实际应用问题，将是接下来的优化方向。考虑到需要处理大量数据并能快速返回检索结果，接下来拟引用企业级搜索引擎 Elastic search 作为技术底座来支持第一步的检索召回研究。③本研究使用的特征主要是基于文本的，比如相似度、共现词、共现词数量，实际上专利中有丰富的字段，比如 IPC、专利引文、专利家族信息，这些都有可能对对比文件查找带来帮助，这些是本研究团队下一步的工作内容。利用人工智能来解决复杂的审

查工作是一项极具挑战性的研究，希望本研究能够为学术界和实务界提供帮助与启示。

参考文献：

[1] 国家知识产权局. 1985 年专利统计年报[EB/OL]. [2020 - 08 - 05]. http://www. cnipa. gov. cn/tjxx/jianbao/1985-1999/85/1.1. htm.

[2] BLOSSER G H, ARSHADI N, AGRAWAL S. A critical assessment of the USPTO policies toward small entity patent applications [J]. Technology and innovation, 2011, 13(3): 249 - 259.

[3] 国家知识产权局. 2018 专利复审无效十大案件[EB/OL]. [2020 - 08 - 05] . http://www. sipo. gov. cn/mtsd/1138630. htm.

[4] 国家知识产权局. 2017 专利复审无效十大案件[EB/OL]. [2020 - 08 - 05]. http://www. sipo. gov. cn/mtsd/1123789. htm.

[5] 国家知识产权局. 申长雨在国家知识产权局专利审查工作座谈会上强调努力提高专利审查质量和效率，推动知识产权事业高质量发展[EB/OL]. [2020 - 08 - 05]. http://www. sipo. gov. cn/zscqgz/1120594. htm.

[6] 国家知识产权局. 2018 年中国知识产权发展状况新闻发布会在京举行[EB/OL]. [2020 - 02 - 05]. http://www. sipo. gov. cn/zscqgz/1138755. htm.

[7] 中华人民共和国国家知识产权局. 专利审查指南(2010)[M]. 北京：知识产权出版社，2009.

[8] 中国专利检索技能大赛[EB/OL]. [2020 - 08 - 05]. http://www. ipsearch. top/home/index. html.

[9] 国家知识产权局专利复审委员会. 以案说法——专利复审、无效典型案例指引[M]. 北京：知识产权出版社，2018：1 - 446.

[10] HUNT D, NGUYEN L, RODGERS M. 专利检索：工具与技巧[M]. 北京市知识产权局，编译. 陈可南，译. 北京：知识产权出版社，2013.

[11] CLARKE N S. The basics of patent searching[J]. World patent information, 2018, 54: S4 - S10.

[12] LUPU M, MAYER K, TAIT J, et al. Current challenges in patent information retrieval[M]. Berlin: Springer, 2011.

[13] 高继刚. 浅析计算机关键词检索的选取在专利检索中的作用[J]. 通讯世界，2015(12)：257 - 257.

[14] 卢士燕，朱佳，李娇，等. 追踪检索在化工领域专利申请审查中的应用[J]. 广东化工，2019，46(3)：131 - 132.

[15] 朱敬敬，杨喆. 专利检索技巧之“顺藤摸瓜”[J]. 科教导刊 - 电子版(上旬)，2017(10)：218 - 220.

[16] 黄微. 专利审查中非专利文献的检索与应用[J]. 中小企业管理与科技(下旬刊)，2016(7)：118 - 119.

[17] RAJSHEKHAR K, SHALABY W, ZADROZNY W. Analytics in post-grant patent review: possibilities and challenges (preliminary report)[J]. Social science electronic publishing, 2017.

[18] 隆瑾. 专利无效对比文件及其获取研究——以专利引文分析为视角[D]. 湘潭：湘潭大学，2012.

[19] 张杰，孙宁宁，张海超. 基于 SAO 结构的中文相似专利识别算法及其应用[J]. 情报学报，2016，35(5)：472 - 482.

[20] 刘玉琴，汪雪锋，吕琳. 基于权利要求结构信息的中文专利无效检索模型[J]. 计算机应用研究，2008，25(7)：2068 - 2070.

[21] 翟东升，马文姗. 中文专利权利要求书分词算法研究[J]. 情报杂志，2011，30(11)：152 - 155.

[22] 马双刚. 基于深度学习理论与方法的中文专利文本自动分类研究[D]. 镇江：江苏大学，2016.

[23] 廖列法，勒孚刚，朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报，2017，37(3)：35 - 39.

[24] 胡杰，李少波，于丽娅，等. 基于卷积神经网络与随机森林算法的专利文本分类模型[J]. 科学技术与工程，2018，18(6)：268 - 272.

[25] GUO M, YUAN H, QIAN Y. A new method for rare feature extraction in patent documents[C]//2016 13th international conference on service systems and service management. Kunming: IEEE, 2016: 687 - 692.

[26] ZHU F, WANG X, ZHU D, et al. User demand-driven patent topic classification using machine learning techniques[C]//The 11th conference on international fuzzy logic and intelligent technologies in nuclear science. Joao Pessoa: World Scientific, 2014: 657 - 663.

[27] CHEN X, DENG N. A semi-supervised machine learning method for chinesse patent effect annotation[C]//2015 international conference on cyber-enabled distributed computing and knowledge discovery. Xi'an: IEEE, 2015: 243 - 250.

[28] KREUCHAUFF F, KORZINOV V. A patent search strategy based on machine learning for the emerging field of service robotics[J]. Scientometrics, 2017, 111(2): 743 - 772.

[29] LEE J. Predicting bad patents[EB/OL]. [2020 - 08 - 05]. https://www2. eecs. berkeley. edu/Pubs/TechRpts/2017/EECS-2017-57. pdf.

[30] WINER D. Predicting bad patents: employing machine learning to predict post-grant review outcomes for US patents [EB/OL]. [2020 - 08 - 05]. https://www2. eecs. berkeley. edu/Pubs/TechRpts/2017/EECS-2017-60. pdf.

[31] HO W. Predicting bad patents[EB/OL]. [2020 - 08 - 05]. https://www2. eecs. berkeley. edu/Pubs/TechRpts/2017/EECS-2017-63. pdf.

[32] YEW T. Predicting bad patents[EB/OL]. [2020 - 08 - 05]. https://www2. eecs. berkeley. edu/Pubs/TechRpts/2017/EECS-2017-66. pdf.

[33] RYAN L, MARCOS T. Predicting patent outcomes with text and attributes[EB/OL]. [2020 - 08 - 05]. http://cs230. stanford. edu/projects_spring_2019/reports/18681598. pdf.

[34] RAJSHEKHAR K, ZADROZNY W, GARAPATI S S. Analytics of patent case rulings: empirical evaluation of models for legal rele-

vance[C]// Proceedings of the 16th international conference on artificial intelligence and law. London: Elsevier,2017:1 – 9.
[35] 邓洁,余翔,崔利刚. 基于专利信息的我国发明专利无效行为实证研究[J]. 情报杂志,2014,33(8):52 – 58.
[36] 李航. 统计学习方法[M]. 北京:清华大学出版社,2019.
[37] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics,2001,29(5):1189 – 1232.
[38] ApacheCN. scikit-learn (sklearn) 官方文档中文版[EB/OL]. [2020 – 08 – 05]. https://sklearn.apachecn.org/.
[39] GENSIM. Core concepts[EB/OL]. [2020 – 08 – 05]. https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html#core-concepts-document.
[40] 万象云. 万象云专利检索[EB/OL]. [2020 – 08 – 05]. https://www.wanxiangyun.net/search/Index.
[41] 中华人民共和国国家知识产权. 专利审查指南(2010)[M]. 北京:知识产权出版社,2010.

作者贡献说明:
郭诗琪:模型调试、数据分析、初稿撰写和论文修改;
贠强:论文选题指导,数据收集、数据处理;
陈亮:提出论文整体研究思路与框架,模型构建、论文修改;
周杰:选题确定,论文修改。

Research on the Method of Judging Reference Document in Patent Invalidation Using GBDT

Guo Shiqi[1,2] Yun Qiang[2] Chen Liang[2] Zhou Jie[2]
[1] Institute of Medical Information/Medical Lirary CAMS&PUMC, Beijing 100020
[2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] Comparative documents are important for judging whether a patent can be granted or invalid. Aiming at the shortcomings of traditional information retrieval methods and rarely using machine learning methods to study the issue of comparative document retrieval, based on the introduction of comparative file information, this paper constructs a patent relevance determination model. [**Method/process**] Experiments were performed by using the target patents and comparative documents in the patent invalidation judgment as the data set to extract text similarity, co-occurrence vocabulary, and co-word quantity feature information. The GBDT model was used to convert the retrieval of comparative documents into classification issues that determined whether they were relevant. [**Result/conclusion**] The research results show that the contribution of different field data to the classification effect is different, in which the F1 of the description text reaches 59%, and the classification effect after multi-feature integration is significantly better than the result of single text similarity. Finally, this paper analyzes the experimental misclassifications and points out the next research directions.

Keywords: patent invalidity the prior art feature selection machine learning